Big data, big opportunities — for science and for enterprise

FEATURE | SEPTEMBER 18, 2013 | BY ANDREW PURCELL

iSGTW speaks to CERN openlab's Sverre Jarp, ahead of this month's inaugural ISC Big Data conference in Heidelberg, Germany. Jarp, who is the general chair of the event, picks his highlights from the conference program and tells us why big data presents exciting opportunities, but also poses challenges, for enterprise and HPC alike.

Perhaps you could start by telling the *iSGTW* readers a little about the event in your own words...

This will be the first-ever 'ISC Big Data' conference and we're looking forward to getting delegates together under the new motto '*where enterprise and HPC meet*' . There's a lot of potential in bringing people together from these two camps and enabling cross-fertilization between enterprise users and HPC/HTC users.

Sverre Jarp speaks about the inaugural ISC Big Data conference, for which he will be general chair. Video courtesy ISC Events. Front page image courtesy CERN openlab.

Why is now the right time for a new conference focusing specifically on big data?

The importance of big data is growing so rapidly. Today, you can hardly open a webpage without discovering somebody talking about big data!

Of course, what's most important is to target this intersection between the experience that is already in the HPC community and the interest that's now coming from the business side. People really want to use data in a much more agile way. We've all heard about the three 'V's of big data (volume, velocity, and variety), so it's no longer just a question of having static data reports on enterprises. They really want to find new ways of crunching the data and getting new business-related information out of it, which can then be used to improve processes, expand markets, drive customer uptake, *etc* .

So, it's fair to say that 'big data' is more than just a buzz phrase then?

It's clearly much more. Over the years, I've been at conferences where people have talked about dealing with gigabytes, terabytes, and now petabytes of data. Each time, there may have been a jump in scale of three orders of magnitude, but people have always wanted this same agility. They want to be able to extract new information that they can plough into their business processes, help them to get a return on their investment, and give them competitive advantages.

What are the highlights of the conference program for you? Are there any talks, workshops, or other sessions, which you are particularly looking forward to?

I think we've put together an outstanding program. I'm really excited about the sheer breadth and depth of the conference. We have speakers from academia, enterprise, and scientific research. We also have a keynote speech planned for each of the two days of the conference. We have another interesting speaker, Michael Feindt from Blue Yonder, who has actually taken analytics algorithms from physics and is applying them to businesses. We even have a speaker from PayPal explaining how the company is now using an HPC approach for fraud detection. It's great to see that with this approach they're able to make a significant return on investment. So, although they have installed 'heavy iron', it's actually well worth it in terms of improving early detection of attempts at fraud.

As well as the opportunities provided by big data, there are also challenges it poses, too. Perhaps you could tell the *iSGTW* readers a little about what you think are the biggest challenges which need to be tackled in order to benefit fully from big data?

Big data is a new paradigm, which means you can no longer just sit in your chair and expect daily sales or transaction reports to come in. With big data, you're really trying to use the data in a much more complex, combinatorial way, so that you can detect new things. It can be problematic because you've got to be sure that your data is correct on the input side, so that you don't have a situation of 'garbage in, garbage out'; you've got to have your complex algorithms ready for deployment; and you've got to know how you're going to use the data to give your business an advantage.

And what about big data specifically in a scientific context? Could you maybe expand upon how big data plays a role in research here at CERN?

CERN has always been at the forefront of what you might call 'physical big data'. We've always been pushing disk storage, as well as tape storage, so that we can cope with as much data as is financially defendable. If you look back, we've always had huge amounts of data: we had terabytes in the days of LEP. Obviously, a terabyte can now be put on one disk and carried away, so clearly time changes things as you look back. Nevertheless, I think that, if you freeze time at any particular point, you'll see that we've always been right up there with the very, very big guys.

Another aspect, of course, is the complex analytics that needs to take place. It's not just a question of running a quick skimming exercise with the physics data and saying 'oh look, we've found the Higgs boson'. There are a lot of complex events which need to be processed and the signal always needs to be separated from the background, too. It is this 'mining' approach which puts CERN firmly on the big data map.

Yet, CERN isn't the only research organization dealing with big data. Members of the EIROforum, which is a partnership between eight of Europe's largest inter-governmental scientific research organizations, have recently collaborated on a report outlining their vision for e-infrastructures in Europe. In this report, they state that 'the era of data-intensive science has begun'.

CERN, EMBL, and ESA are also collaborating with leading IT providers, from both the public and private sectors, on the Helix Nebula initiative, which aims to create a cloud computing infrastructure to support the massive IT requirements of European scientists. Rupert Lueck will be giving a talk at ISC Big Data '13 explaining how EMBL, as one of Helix Nebula's three flagship use cases, is able to implement a novel cloud service to simplify large-scale genome analysis and to help scientists, inside and outside EMBL, to better meet the challenges of analyzing large amounts of genomic sequence data by provisioning tailor-made high performance computing and bioinformatics resources on demand.

As the chief technology officer for CERN openlab, you clearly have much experience of working at the interface between public and private research. This seems to tie in nicely with the motto for ISC Big Data '13, which you mentioned at the start of this interview. How do you personally think big data is bringing enterprise and HPC together then?

CERN is a scientific institute, but if you were blindfolded and taken into our data center, it would be hard for you to say whether this data center is actually part of a research institute, or belongs to a business enterprise, such as, say, UPS in Basel, for instance. The data centers would, to a large extent, be based on the same components, have the same vendors, and a lot of the same software packages applied, *etc.*

IT companies tend to make all of their products homogenous. So, whether it's applied to science or enterprise, you're really dealing with the same components. Maybe they're configured in different ways, but boundaries are blurred. For as long as I've been at CERN, there have always been many commonalities. Even back when we used to have mainframes, we could go to conferences and talk to people from the big banks, insurance companies, and so on, and they too would also be running mainframes. The deployment, in terms of the details, may have been slightly different, but the offerings from the companies producing the mainframes were, of course, the same.

As such, there are many commonalities between the partners in CERN openlab, with vendors and customers easily able to talk to one another without feeling that they are operating in completely different worlds. I would really like to see the scientific and enterprise communities walk hand in hand, especially in terms of collaborating jointly with vendors, so that future hardware and software developments become something that both worlds can use very effectively.

ISC Big Data will be held on 25-26 September in Heidelberg, Germany. Read more about the conference on the event website, here.

Also, find out more about the upcoming ISC Cloud '13 conference, to be held in Heidelberg on 23-24 September, here.

Average:

Your rating: None Average: 4.3 (3 votes)

About the Author »

Andrew Purcell

Editor

Andrew Purcell is the editor of iSGTW and is based at CERN, near Geneva.

RELATED TERMS: big data CERN CERN openlab Europe grid computing Heidelberg High-performance computing high-throughput computing HPC HTC ISC ISC Big Data ISC Cloud ISC events ISC'13 openlab Sverre Jarp cloud computing data management systems workflow management systems physics and astronomy

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

By submitting this form, you accept the Mollom privacy policy.

SAVE PREVIEW